УДК 004.6

СПЕЦИФИКА ОБУЧЕНИЯ СТУДЕНТОВ-МАТЕМАТИКОВ СПЕЦИАЛИЗАЦИИ “НАУКА О ДАННЫХ” В ПРОГРАММАХ АКАДЕМИЧЕСКОГО ВЫСШЕГО ОБРАЗОВАНИЯ

Главацкий С.Т., кандидат физико-математических наук, доцент,
Московский государственный университет имени М.В.Ломоносова, г. Москва
serge@rector.msu.ru

Бурыкин И.Г., научный сотрудник,
Московский государственный университет имени М.В.Ломоносова, г. Москва
ilia.burykin@sdo.msu.ru

Аннотация. В статье изложен авторский взгляд на формирование и преподавание специализации “Наука о данных” для математиков. Подчеркивается ориентация на фундаментальность подхода к построению (математических) моделей данных, строгой постановке задач исследования и разработке методов и алгоритмов их решения. Предлагается набор специальных и общих курсов, необходимых для подготовки специалиста по данным.

Ключевые слова: большие наборы данных, анализ данных, интеллектуальный анализ данных, извлечение новых знаний, типы данных, структуры данных, модели данных, машинное обучение, искусственный интеллект.

SPECIFICS OF TEACHING STUDENTS-MATHEMATICIANS THE SPECIALIZATION “DATA SCIENCE” IN THE PROGRAMS OF ACADEMIC HIGHER EDUCATION

S.T. Glavatsky, candidate of physico-mathematical sciences, associate professor,
Moscow Lomonosov State University, Moscow
serge@rector.msu.ru

I.G. Burykin, researcher,
Moscow Lomonosov State University, Moscow
ilia.burykin@sdo.msu.ru

Abstract. The article presents an author's view on the formation and teaching of the specialization "Data Science" for mathematicians. An emphasis is placed on the fundamental nature of the approach to constructing (mathematical) data models, rigorous formulation of research problems and the design of methods and algorithms for their solutions. A set of special and general courses is offered to prepare a data specialist.

Keywords: Big Data, data analysis, data mining, new knowledge extraction, data types, data structures, data models, machine learning, artificial intelligence.

В настоящее время происходит взрывной рост технологических решений и научных исследований в области больших наборов данных (“Big Data”). В мировом сообществе сложилось представление о больших наборах данных, как характеризующихся следующими основными особенностями [1]:

- объемом (Volume);
- скоростью обновления (Velocity);
- разнообразием и неоднородностью (Variety);
- проблемами с достоверностью (Veracity);
- стоимостью обработки (Value);
- изменчивостью (Variability);
- потребностью в визуализации (Visualization).

В последние 2-3 года “большие данные” из экспериментальных новых технологий переросли в основные корпоративные системы, фактически развернутые в производстве. И это, в свою очередь, вызвало потребность в специалистах и исследователях, умеющих работать с “большими данными”. Специализация “учёного по данным” или, другими словами, “специалиста по работе с [большими] данными” (“data scientist”) сейчас считается одной из самых привлекательных, высокооплачиваемых и перспективных профессий. Подготовка таких специалистов сейчас, в основном, происходит в рамках специализации называемой “Наука о данных” (“Data Science”).

“Наука о данных” или “даталогия” (“Datalogy”), начиная с 70-х годов прошлого века, рассматривается как академическая дисциплина, а с начала 2010-х годов, во многом благодаря популяризации концепции “больших данных”, – и как практическая межотраслевая сфера деятельности.

Существует множество подходов и точек зрения на содержание “науки о данных” и её месте в области прикладной математики и информатики. Например, “науку о данных” можно классифицировать как раздел информатики, изучающий проблемы [2]:

- анализа данных;
- обработки данных;
- представления данных в цифровой форме.

“Наука о данных” объединяет:

- методы по обработке данных в условиях больших объёмов и высокого уровня параллелизма;
- статистические методы;
- методы интеллектуального анализа данных и приложения искусственного интеллекта для работы с данными;
- методы проектирования и разработки баз данных (БД).

Можно выделить следующие основные научно-инженерные направления “науки о данных”:

- большие наборы данных, т.е. сбор и обработка больших объемов данных;
- анализ данных и построение средств поддержки принятия решений;
- разработка и использование статистических и математических моделей, алгоритмов и визуализаций;
- интеллектуальный анализ данных, извлечение новых знаний;
- бизнес-аналитика;
- эконометрика;
- статистика;
- машинное обучение;
- искусственный интеллект;
- математическое моделирование.

Обычно, преподавание “науки о данных” содержит учебные планы по нескольким дисциплинам и осуществляется в рамках таких устоявшихся жизненно важных областей, как [3]:

- информатика / Computer science (CS2013);
- компьютерная инженерия / Computer engineering (CE2016);
- информационные системы / Information systems (IS2010);
- информационные технологии / Information technology (IT2017 in progress);
- разработка программного обеспечения / Software engineering (SE2014);
- кибербезопасность / Cybersecurity (CSEC2017 in progress).

Причем каждая область обладает своей собственной идентичностью и педагогическими традициями. Интересно также отметить, что помимо вышеперечисленных областей, сейчас происходит разработка учебных планов непосредственно для “науки о данных”.

Проблемы обучения студентов “науке о данных” и подготовки “учёного по данным” имеют свою специфику для классических университетов, готовящих специалистов в рамках программ академического высшего образования.

На кафедре теоретической информатики механико-математического факультета МГУ имени М.В. Ломоносова в течение ряда лет разрабатывается цикл специальных курсов и практикумов под

общим наименованием “Аналитика больших данных для математиков” (“Data Science and Data Mining for Mathematicians”) [4]. В рамках этого направления преподавания на базе имеющихся общих курсов, а также за счет введения новых общих дисциплин, развивается преподавание отдельной специализации по методам и алгоритмам представления, моделирования и анализа больших наборов данных.

Нашей целью в преподавании “Науки о данных” для математиков является, в определенном смысле, выделение подходов к исследованиям, основанных на:

- использовании математических теорий, понятий и моделей;
- постановке и решении математических задач;
- применении разработанных алгоритмов в решении задач обработки и анализа данных.

Мы предлагаем студентам и аспирантам следующие учебные курсы:

i. Модели данных и базы данных (Data models and databases) – годовой курс по выбору кафедры, включающий в себя:

1. Модели данных и основы систем баз данных (Data models and fundamentals of database systems) – полугодовой курс по выбору кафедры;
2. Базы данных: дополнительные главы (Databases: additional chapters) – полугодовой курс по выбору студента;

ii. Аналитика больших данных (Big Data Analytics) – годовой курс по выбору кафедры, включающий в себя:

1. Аналитика больших данных: основные алгоритмы (Big Data Analytics: basic algorithms) – полугодовой курс по выбору кафедры;
2. Аналитика больших данных: дополнительные главы (Big Data Analytics: additional chapters) – полугодовой курс по выбору студента.

Эти курсы:

- имеют теоретическую и практическую составляющие;
- являются, с одной стороны, взаимозависимыми, а с другой – не требуют обязательного предварительного изучения содержания остальных спецкурсов из предлагаемого набора;
- отражают как уже ставшие классическими модели и алгоритмы, так и современные взгляды и понятия.

Основными темами, изучаемыми в этих курсах, являются:

- типы данных, структуры данных, модели данных;
- представление данных, хранение и передача данных;
- методы и алгоритмы первичной обработки данных, базы данных, языки манипулирования данными;
- проектирование баз данных, языки определения данных, нормальные формы в проектировании реляционных баз данных;
- структурированные и неструктурированные данные, хранилища данных;
- анализ больших (неструктурированных) наборов данных, технологии распараллеливания обработки и сжатия информации;
- задачи интеллектуального анализа больших наборов данных и проблемы больших объемов и размерностей;
- вероятностные методы первичного сжатия данных, хеширование и статистические оценки;
- задача обнаружения схожих документов, предлагаемые методы и алгоритмы, применение технологий распараллеливания обработки;
- метрические пространства, кластерные методы в снижении размерности задачи;
- рекомендательные системы, матричное представление данных, алгоритмы линейной алгебры и их использование в снижении размерности задачи;
- “всемирная паутина” (WWW), методы сбора данных и первичного анализа;
- структура “всемирной паутины” и ее использование в задачах ранжирования информации;
- интеллектуальный анализ информационных процессов;

- продвинутые техники баз данных, In-Memory базы данных как технологическая платформа для обработки больших наборов данных;
- базы данных NoSQL как набор технологических платформ для обработки больших наборов данных.

Предполагается, что слушатели курсов уже владеют материалом из основных курсов по:

- линейной алгебре и ее приложениям;
- по теории вероятностей и статистике;
- по программированию;
- по теории кодирования.

При этом, по нашему мнению, знание линейной алгебры слушателями является гарантией успешного прохождения обучения. Современными исследователями высказывается мнение (Skyler Speakman), что “Линейная алгебра есть математика XXI века”. С одной стороны удивительно, а с другой стороны – очень интересно наблюдать, как математическая дисциплина, фактически полностью сформировавшаяся к середине XIX века, становится весьма актуальной как в начале XX века, так и сейчас – в начале XXI столетия. Актуальной, по крайней мере, для “учёного по данным”, поскольку для понимания методов анализа “больших данных” необходимо знание таких тем как:

- сингулярное разложение матриц (SVD);
- собственное разложение матриц, главные компоненты;
- LU-разложение матриц;
- QR-разложение / факторизация матриц;
- симметричные матрицы;
- ортогонализация и ортонормализация;
- матричные операции;
- проекции;
- собственные значения и собственные векторы;
- векторные пространства и нормы.

В качестве фундаментальных основ науки о данных студентам в обязательном порядке предлагаются также темы о структурированных данных:

- модели данных, реляционная модель, реляционная алгебра, основные операторы, свойства, запись операторов в линейной и древовидной форме;
- реляционная СУБД, язык SQL, структура, команды, выразимость, реализация основных функций ACID;
- проектирование схем БД, функциональные зависимости, реализация алгоритмов замыкания множеств атрибутов и множеств зависимостей;
- нормальные формы схем отношений, реализация алгоритмов декомпозиции в 3НФ и нормальную форму Бойса-Кодда;
- концептуальное моделирование, ER-модель, преобразование в реляционную модель данных;
- административное управление базами данных, преобразование схем БД, проверка целостности данных, восстановление данных.

Для успешного восприятия материала курсов студентам предлагается не только теоретическая часть, но и её практическая поддержка в виде выполнения конкретных проектов с использованием предустановленных программных сред, в частности,

- SAP SQL Anywhere 17 Developer Edition и
- SAP HANA, Express Edition 2.0 SPS02 (Virtual Machine Method).

Для работы с помощью интерактивной аналитики с источниками "больших данных" в Hadoop-ландшафте (Apache Spark) и для изучения языка Spark SQL студент может воспользоваться SAP Vora 1.4 Developer Edition.

В заключении отметим, что сейчас мы являемся свидетелями появления нового тренда – “большие данные + искусственный интеллект”, в котором:

- технологии “больших данных” используются для решения основных задач обработки данных;

- “машинное обучение” используется для извлечения новых знаний из данных (в виде аналитических идей или действий).

Поэтому мы рассматриваем вопрос о включении элементов “машинного обучения” (“machine learning” / “deep learning”) в линейку наших курсов.

Литература

1. Livingstone R. The 7 Vs of Big Data. [Электронный ресурс] / Livingstone R. – Режим доступа: http://rob-livingstone.com/2013/06/big-data-or-black-hole
2. Наука о данных. [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/
3. Сухомлин В.А. Международные образовательные стандарты в области информационных технологий / Сухомлин В.А. // Прикладная информатика. – 2012. – № 1 (37). – С. 33–54.
4. Glavatsky S. About courses cycle "Data science and data mining for mathematicians / Glavatsky S., Burykin I. // CEUR Workshop Proceedings (CEUR-WS.org): Selected Papers of the XI International Scientific-Practical Conference Modern Information Technologies and IT-Education (SITITO 2016), Moscow, Russia, November 25-26, 2016. – Vol. 1761. – 2016. – P. 58–63.